

## TREATMENT EVALUATION IN THE PRESENCE OF SAMPLE SELECTION

**Martin Huber**

*Department of Economics, University of St. Gallen, St. Gallen, Switzerland*

□ *Sample selection and attrition are inherent in a range of treatment evaluation problems such as the estimation of the returns to schooling or training. Conventional estimators tackling selection bias typically rely on restrictive functional form assumptions that are unlikely to hold in reality. This paper shows identification of average and quantile treatment effects in the presence of the double selection problem into (i) a selective subpopulation (e.g., working—selection on unobservables) and (ii) a binary treatment (e.g., training—selection on observables) based on weighting observations by the inverse of a nested propensity score that characterizes either selection probability. Weighting estimators based on parametric propensity score models are applied to female labor market data to estimate the returns to education.*

**Keywords** Inverse probability weighting; Sample selection; Treatment effects.

**JEL Classification** C13; C14; C21.

### 1. INTRODUCTION

The sample selection problem, which was discussed by Gronau (1974), Heckman (1974), and Vella (1998), among many others, arises whenever the outcome of interest is only observable for some subpopulation that is non-randomly selected even conditional on observed factors. Potential bias due to sample selection related to unobserved characteristics is an issue for a range of treatment evaluation problems, e.g., when estimating the returns to schooling based on a selective subpopulation of working or the effect of school vouchers on college admissions tests, given that some students abstain from the test in a non-random manner.

This paper discusses treatment evaluation under sample selection and attrition when the treatment assignment is non-random and related to

observed factors. It considers the case of a double selection problem into (i) the subpopulation for which the outcome is observed (selection on unobservables) and (ii) the treatment (selection on observables). The main contribution is to show that average and quantile treatment effects are identified by weighting observations by the inverse of a nested propensity score which controls for sample selection bias in the subpopulation with observed outcomes (e.g., working) and treatment selection bias due to non-random treatment assignment.

The present work is related to the literature on inverse probability weighting (IPW), which has long been known as a general approach to tackle selection problems, see Horvitz and Thompson (1952). In the literature on missing data, attrition, and sample selection, Robins and Rotnitzky (1995), Robins et al. (1995), Rotnitzky and Robins (1995), and Wooldridge (2002, 2007) weight regressions by the inverse of the sample selection propensity score, i.e., the conditional probability to be observed. However, they do not consider selection on unobservables as in this paper. In the treatment evaluation literature relying on the selection on observables or conditional independence assumption (CIA) (see for instance Imbens, 2004), Hirano et al. (2003) and Firpo (2007) study IPW estimators of average and quantile treatment effects based on weighting by the inverse of the treatment propensity score, the conditional probability to be treated, to control for selection into treatment. Bang and Robins (2005) use IPW in regression models separately for sample selection and treatment selection problems. This paper adds to the literature on IPW by considering both problems within the same model. Identification of treatment effects relies on the inclusion of the (first stage) sample selection propensity score, which is identified using an exclusion restriction, as additional covariate among other observed factors in the (second stage) treatment propensity score.

The paper also contributes to the classic sample selection literature. Under nonparametric identification of the sample selection and treatment propensity scores the identification of treatment effects based on IPW is nonparametric, too. This framework invokes weaker restrictions than the fully parametric selection model in Heckman (1974, 1976, 1979). It is also more general than the semiparametric models of Ahn and Powell (1993), considering a nonparametric sample selection process (e.g., the decision to work), and Newey (2009), considering semiparametric sample selection, who, however, all impose linearity in the outcome equation. Therefore, the selection model discussed in this paper allows for heterogenous effects with respect to observed factors such that the effects may be different for different populations. For this reason the next section discusses identification for various target populations that appear to be interesting for policy interventions. Finally, our model is slightly more general than that of Das et al. (2003), who consider a nonparametric

sample selection model but still impose additivity of the unobservables which need not be assumed here. Under a parametric specification of the nested propensity score (as considered in the empirical application), identification of treatment effects is semiparametric. This framework is more restrictive with respect to the sample selection process than Ahn and Powell (1993) and Newey (2009), but more general with respect to effect heterogeneity.

As in the classic sample selection literature, an exclusion restriction is used to identify the sample selection propensity score. Endogeneity only emerges from the sample selection problem. This is distinct from the instrumental variable (IV) literature considering endogenous treatments, see for instance Imbens and Angrist (1994) and Frölich and Melly (2008). Identification in this paper is based on an instrument for sample selection, whereas the IV literature instruments the endogenous treatment directly. Which of the two approaches is accurate, if any, depends on the evaluation problem, the target population, and the data at hand. The framework considered is for example also different to the empirical application in Ahn and Powell (1993), where sample selection and endogeneity in regressors of the outcome equation arises in the same evaluation problem. This requires distinct instruments for selection and the endogenous regressors, whereas we assume conditional exogeneity of the treatment and only instrument selection.

Estimators of the average treatment effect (ATE) and quantile treatment effect (QTEs) naturally arise from the sample analogues of the identification results. Alternatively to IPW estimation, matching estimators (see Rubin, 1973a,b, 1976) on the nested propensity score can be used. Given the importance of semiparametric estimation in the empirical treatment evaluation literature, we apply semiparametric IPW and matching (using probit models for the propensity score specifications) to a repeated cross-section (1975–1979) from the U.S. Current Population Survey (CPS) previously analyzed by Mulligan and Rubinstein (2008). We estimate the wage differentials between females who went to high school with and without graduation and find that graduating increases average weekly wages by roughly 17% over dropping out of high school. Furthermore, the graduation effects appear to be larger at higher ranks of the wage distribution. As a robustness check, we also estimate bounds by invoking assumptions previously used by Lechner and Melly (2007), Lee (2009), Zhang and Rubin (2003), and Zhang et al. (2008).

The remainder of the paper is organized as follows. Section 2 introduces a general sample selection model and discusses identification of average and quantile treatment effects for various populations of interest. Section 3 briefly discusses estimation based on IPW, which proceeds in three steps. An empirical application of IPW, propensity score matching (PSM), and the estimation of bounds to labor market data from the CPS

is presented in Section 4. Section 5 provides simulation results about the finite sample properties of the IPW and PSM estimators. Section 7 concludes.

## 2. MODEL AND IDENTIFICATION

### 2.1. Basic Notation and Setup

Before going into the details of our model and the identification strategy, we briefly summarize the notation to be used in Table 1 to facilitate later reference.

We are interested in the effect of  $D$  on  $Y$  but face the problems that the assignment to  $D$  is selective and that  $Y$  is only observed conditional on  $S = 1$ . For this reason, our identification strategy will rely on assumptions related to the observed covariates  $X$  and the instrument for selection  $Z$ , as elaborated in the next sections.

### 2.2. Model

We introduce a general sample selection model, where the latent outcome is an unknown function of two observed components, the treatment of interest and a vector of covariates, and an unobserved term.  $Y$  denotes the latent outcome that is only partially observed conditional on selection, represented by the binary variable  $S$ . Let  $D$  denote a treatment, which is either 1 (treatment) or 0 (non-treatment). Even though the subsequent discussion focusses on the binary treatment case, it could be easily extended to multiple treatments as discussed in Imbens (2000) and Lechner (2001). Let  $X$  and  $U$  denote the covariates and the unobserved term, respectively. Throughout the paper, we will assume to have an independent and identically distributed (i.i.d.) sample of  $n$  units before

**TABLE 1** Notation

Symbol	Meaning
$Y$	Outcome variable (discrete or continuous), only observed conditional on selection
$S$	Binary selection indicator
$D$	Binary treatment variable, the effect of which is of interest
$X$	Observed covariates, may affect both the outcome and selection
$U$	Unobserved factor affecting the outcome
$V$	Unobserved factor affecting selection, may be related with $U$
$Z$	Instrument affecting selection, independent of the unobserved factors
$W$	Shortcut notation for $(D, X, Z)$
$p(W)$	Selection propensity score, defined as $\Pr(S = 1   D, X, Z)$
$\pi(X, p(W))$	Treatment propensity score, defined as $\Pr(D = 1   X, p(W))$
$Y^1, Y^0$	Potential outcomes when setting treatment $D$ equal to 1 or 0

sample selection takes place, indexed by  $i = 1, \dots, n$ . For the latent outcome, we assume the model

$$Y_i = \varphi(D_i, X_i, U_i), \quad (1)$$

where  $\varphi(\cdot)$  is an unknown function.

We observe  $\{X_i, D_i\}$  for all units in the sample, but the outcome  $Y_i$  only conditional on  $S_i = 1$ . Empirical examples for such setups include wage regressions (where  $S$  is employment), see Gronau (1974) and Heckman (1974, 1976), or the evaluation of the effects of policy interventions in education on test scores (where  $S$  is participation in the test), see Angrist et al. (2006) and Angrist et al. (2009). The selection indicator  $S$  is assumed to be a function of the treatment, the covariates, an instrument, and an unobserved term:

$$S_i = I\{\zeta(D_i, X_i, Z_i) \geq V_i\}. \quad (2)$$

$I\{\cdot\}$  denotes the indicator function, and  $\zeta(\cdot)$  is an unknown function.  $Z$  represents a one- or multidimensional instrument which is observable for all units and not directly related with the outcome.  $V$  is an unobserved term that is possibly related with  $U$ . Due to the dependence of  $V$  and  $U$ , the observed outcomes are a non-random subsample of latent outcomes. By assumption,  $S$  is a function of one element that is excluded in  $\varphi$ , namely the instrument  $Z$ . Point identification of treatment effects crucially hinges on this exclusion restriction.  $Z$  has to be relevant for  $S$  in the sense that it shifts the selection probability considerably conditional on  $D, X$ , and in general, at least one element of the instrument needs to be continuous.

Models alike the one defined by Eqs. (1) and (2) have been referred to as models with partial observability, see for instance Poirier (1980) and Meng and Schmidt (1985) who consider fully parametric specifications in which both the selection and outcome variables are binary. A prominent economic problem to which our more general model may be applied are the returns to schooling or training. In this case,  $Y$  denotes the potential wages which are only observed conditional on employment ( $S = 1$ ), and  $D$  represents participation in a training program or educational attainment.  $X$  includes other factors that determine wages and are possibly related with  $D$  such as work experience. The sample selection problem arises if unobserved factors as motivation affect both the employment decision and potential wages. Identification therefore requires at least one variable ( $Z$ ) that is related with the employment decision but has no direct effect on wages. In the empirical literature on female wage equations, the number of small children in the household and non-wife income have been frequently used as instruments.

### 2.3. Identification

To identify the causal effects of  $D$ , we utilize the potential outcome framework advocated by Rubin (1974), among others. We denote the potential outcome for individual  $i$  and some hypothetical treatment  $D = d$  as

$$Y_i^d = \varphi(d, X_i, U_i).$$

The difference  $Y_i^1 - Y_i^0$  would identify the individual treatment effect, but is unknown to the researcher, because each individual is either treated or not treated and cannot appear in both states of the world at the same time. As an additional complication, the outcomes are observed for a selective subpopulation. Therefore, effects are only identified when further assumptions are invoked.

If treatment effects were homogenous as assumed in the classic sample selection literature (e.g., Heckman, 1974, 1976, 1979), they would be equal for any individual and population, but this seems implausible for most evaluation problems. Therefore, treatment effects are most likely different for different populations considered. Which target population is most interesting from a policy perspective depends on the particular problem at hand. Lee (2009) and Zhang et al. (2008) consider treatment effects on the subpopulation that is always selected irrespective of the treatment assignment, whereas Lechner and Melly (2007) focus on the subpopulation that is selected and treated. In the subsequent discussion, we will first identify the treatment effects on the subpopulation with observed outcomes, i.e., conditional on being selected, and then show identification for the total population by imposing somewhat stronger conditions. After having established our main results, we will also discuss how effects on further target populations can be identified.

For the moment, let us assume that we want to learn about the ATE, denoted as  $\Delta_{S=1}$ , and QTE, denoted as  $\Delta_{S=1}^\tau$ , on the subpopulation with observed outcomes:

$$\Delta_{S=1} = E[Y^1 | S = 1] - E[Y^0 | S = 1],$$

$$\Delta_{S=1}^\tau = Q_{Y^1 | S=1}^\tau - Q_{Y^0 | S=1}^\tau.$$

$\tau$  denotes the rank of the potential outcome distribution at which the QTE is evaluated and is bounded between 0 and 1. For example,  $\tau = 0.5$  yields the median effect of the treatment.  $Q_{Y^d | S=1}^\tau$  denotes the quantile of the potential outcome for treatment  $D = d$  in the subpopulation with observed outcomes and is defined as  $\inf_y \Pr(Y^d \leq y | S = 1) \geq \tau$ .

Briefly speaking, identification in this paper is based on 3 key assumptions: (i) the conditional independence of potential outcomes and

treatments in the total population, (ii) the availability of an exclusion restriction to identify the sample selection propensity score, and (iii) the conditional independence of observables and unobservables given the sample selection propensity score.

**Assumption 1.** *Conditional independence of treatments and latent potential outcomes.*

- (1a)  $Y^1, Y^0 \perp D \mid X = x, \forall x$  in the support of  $X$  (conditional independence of the latent outcome).  
 (1b)  $0 < \Pr(D = 1 \mid X = x) < 1, \forall x$  in the support of  $X$  (common support of  $D$  in  $X$ ).

The CIA or selection on observables assumption is frequently imposed in the treatment evaluation literature, see for instance Heckman et al. (1997) and Lechner (1999). (1a) states that the potential latent outcome is independent of the treatment given the observed covariates  $X$ . This implies that all factors jointly affecting the treatment assignment and the latent outcome can be controlled for by conditioning on the covariates. The difference to conventional evaluation studies relying on the CIA is that the outcome is not fully observed. (1b) is a common support assumption and states that the selection probability must not be perfectly predicted conditional on the covariates. If in addition to Assumption 1, the Stable Unit Treatment Value Assumption (SUTVA) (see Rubin, 1990) is satisfied, stating that the potential outcome for any individual is stable in the sense that it takes the same value independent of treatment allocations in the rest of the population, it holds that

$$\begin{aligned} E[Y^1 \mid D = 0, X = x] &= E[Y^1 \mid D = 1, X = x] = E[Y \mid D = 1, X = x], \\ E[Y^0 \mid D = 1, X = x] &= E[Y^0 \mid D = 0, X = x] = E[Y \mid D = 0, X = x]. \end{aligned}$$

The ATE conditional on  $X$  is  $\Delta(x) = E[Y^1 \mid X = x] - E[Y^0 \mid X = x] = E[Y \mid D = 1, X = x] - E[Y \mid D = 0, X = x]$ . Thus, under Assumption 1, the effect of  $D$  on  $Y$  *could* be identified conditional on  $X$  if the outcome *was* fully observed. However, as unobservables  $V$  and  $U$  are not independent even conditional on  $X$ , the treatment effect is confounded in the subpopulation with observed outcomes. Point identification requires the availability of an instrument  $Z$  that predicts selection  $S$  but is not related with  $Y$  conditional on  $D, X$ . We therefore make the following assumption.

**Assumption 2.** *Exclusion restriction.*

- (2a)  $\text{Cov}(Z, S \mid X, D) \neq 0$  and  $Y \perp Z \mid D, X$  (exclusion restriction).

- (2b)  $\Pr(S = 1 \mid D = d) > c$ ,  $c > 0$ ,  $d \in \{1, 0\}$  (positive conditional selection probability given  $D$ ).
- (2c)  $(U, V) \perp (D, Z) \mid X$ ,  $\Pr(S = 1 \mid D, X, Z)$  (conditional independence of unobservables and  $D, Z$  given  $X$ ).
- (2d)  $F_V(t)$ , the cdf of  $V$ , is strictly monotonic in the argument  $t$ .

Assumption (2a) states that  $Z$  shifts  $S$  but is independent of the latent outcome given  $D, X$ . Direct effects of  $Z$  on  $Y$  are ruled out. Together with Assumption 1, this implies that  $F_{Y \mid D, X}$ , the conditional cdf of  $Y$  given  $D, X$ , is equal to  $F_{Y \mid D, X, Z}$ , the conditional cdf given  $D, X, Z$ , for all values of  $Z$ . (2b) rules out that being treated or nontreated perfectly predicts nonselection. Here,  $c$  represents any positive constant by which  $\Pr(S = 1 \mid D = d)$  is bounded away from zero. We will use  $c$  in other assumptions presented further below, even though its value need not be the same across the different assumptions. To see the usefulness of (2b), assume the opposite, that units with  $D = 0$  are never selected independent of the values of  $X, Z$ . Obviously, the treatment effect cannot be evaluated as no comparisons with  $D = 0$  are available in the subpopulation with observed outcomes.

By (2c), we impose that  $D, Z$  are jointly independent of the unobservables  $U, V$  given  $X$  and the conditional selection probability  $\Pr(S = 1 \mid D, X, Z)$ . (2c) is for instance violated if  $U$  is related to  $D$  in the total population conditional on  $X$  (and  $\Pr(S = 1 \mid D, X, Z)$  which will be kept implicit in the subsequent discussion). Then, the selection bias cannot be controlled for by controlling for  $X$ , as unobserved interaction terms of  $U$  and  $D$  drive the selection probability. To illustrate this issue by means of an example, assume that we are interested in the effects of a training ( $D$ ) on wages ( $Y$ ) and that motivation ( $U$ ) is not observed. Assumption (2c) would be violated if the variance of motivation (and thus, of potential wages) differed for individuals with and without training, but with the same observed factors like age, education, work experience, and others. Albeit strong, equivalent or similar assumptions are crucial for point identification in any selection model of both parametric and general form.

Note that  $\Pr(S = 1 \mid D, X, Z) = \Pr(\zeta(D, X, Z) \geq V) = F_V(\zeta(D, X, Z))$ . By the monotonicity assumption (2d), it holds that the likelihood to be selected increases monotonically in  $\zeta$ . Monotonicity is implicitly assumed in any linear index restriction frequently used in the sample selection literature. However, it is a rather strong restriction, and its plausibility needs to be evaluated from case to case. For example, if  $V$  reflects ability or motivation and  $S$  is employment, it seems reasonable to assume that (2d) holds, as more able and motivated individuals may have a higher intrinsic utility from work and also higher potential wages (extrinsic utility). As a second example, let  $S$  denote summer school participation



and  $V$  ability. If the least able students are likely to participate due to force and the most able students due to personal interest, the monotonicity assumption clearly fails.

By comparing individuals with the same response propensity score under the satisfaction of (2d), we control for  $V$  and thus, also for the dependence between  $V$  and  $U$ . That is, by fixing  $V$ , we rule out confounding of the treatment effect due to attrition related to unobservables. The response propensity score serves as a control function where the exogenous variation comes from  $Z$ . Control functions have been applied in semi- and nonparametric sample selection models, e.g., Ahn and Powell (1993) and Das et al. (2003) as well as in nonparametric models with endogeneity, see for example Newey et al. (1999), Blundell and Powell (2003), and Imbens and Newey (2009).

For notational ease, let  $W \equiv (D, X, Z)$  and  $p(W) \equiv \Pr(S = 1 | D, X, Z)$ . Under Assumption 2,  $U$  and  $D$  are independent conditional on  $p(W)$  and  $X$ , which can be shown analogously to the proof of Theorem 1 in Newey (2007). Let  $a(U)$  denote any bounded function of  $U$ . Note that the incidence of  $S = 1$  can be equivalently expressed as  $F_V^{-1}(p(W)) \geq V$ . Then,

$$\begin{aligned} E[a(U) | D, X, p(W), S = 1] &= E[E[a(U) | V, D, X, Z] | D, X, p(W), F_V^{-1}(p(W)) \geq V] \\ &= E[E[a(U) | V, X] | D, X, p(W), F_V^{-1}(p(W)) \geq V] \\ &= E[E[a(U) | V, X] | X, p(W), F_V^{-1}(p(W)) \geq V] \\ &= E[E[a(U) | V, X, p(W)] | X, p(W), S = 1] \\ &= E[a(U) | X, p(W), S = 1], \end{aligned}$$

where the first equality follows from iterated expectations, the second and third from (2c), and the last from a backward application of the law of iterated expectations.

Thus, as any bounded function of  $U$  and  $D$  are independent conditional on  $p(W)$  and  $X$ , sample selection bias among those with observed outcomes can be controlled for by including the sample selection propensity score as additional conditioning variable besides the covariates  $X$ . To see this, note that the conditional ATE given  $X$  and  $p(W)$  in the selected subpopulation is defined as

$$\begin{aligned} \Delta_{S=1}(x, p(w)) &= \int \varphi(1, x, u) dF_{u | X=x, p(W)=p(w), S=1} \\ &\quad - \int \varphi(0, x, u) dF_{u | X=x, p(W)=p(w), S=1} \end{aligned}$$

$$\begin{aligned}
&= E[Y^1 | X = x, p(W) = p(w), S = 1] \\
&\quad - E[Y^0 | X = x, p(W) = p(w), S = 1].
\end{aligned}$$

$E[Y^d | X = x, p(W) = p(w), S = 1]$  is the expected potential outcome for a hypothetical treatment  $d$  given  $X$  and  $p(W)$  in the subpopulation with observed outcomes. By the conditional independence of  $U$  and  $D$  given  $p(W)$  and  $X$ , it holds that

$$\begin{aligned}
&E[Y^d | X = x, p(W) = p(w), S = 1] \\
&= \int \varphi(d, x, u) dF_{u | X=x, p(W)=p(w), S=1} \\
&= \int \varphi(d, x, u) dF_{u | D=d, X=x, p(W)=p(w), S=1} \\
&= E[Y | D = d, X = x, p(W) = p(w), S = 1].
\end{aligned}$$

Hence, the expected *potential* outcome is equal to the expected *conditional* outcome given  $D = d$ . The ATE  $\Delta_{S=1}$  is identified by the integration over the marginal distributions of  $X$  and  $p(W)$  in the subpopulation with observed outcomes:

$$\begin{aligned}
&\int \int [E[Y | D = 1, X = x, p(W) = p(w), S = 1] \\
&\quad - E[Y | D = 0, X = x, p(W) = p(w), S = 1]] dF_{x | p(W)=p(w), S=1} dF_{p(w) | S=1} \\
&= \int \int [E[Y^1 | X = x, p(W) = p(w), S = 1] \\
&\quad - E[Y^0 | X = x, p(W) = p(w), S = 1]] dF_{x | p(W)=p(w), S=1} dF_{p(w) | S=1} \\
&= E[Y^1 - Y^0 | S = 1] = \Delta_{S=1}.
\end{aligned} \tag{3}$$

In contrast to the ATE, the identification of QTEs requires that the outcome variable is continuous and that the conditional quantiles of interest are unique; that is, the density in the neighborhood of the quantiles must be bounded away from zero such that each quantile corresponds to exactly one particular rank in the conditional distribution. Furthermore, for an intuitive interpretation of QTEs, the rank stability assumption has to be satisfied across treatments. It states that individuals occupy the same rank in potential outcome distributions for different treatments, see for instance Firpo (2007) for more discussion.

Let  $Q_{Y^d | S=1}^\tau(x, p(w))$  denote the  $\tau$ th quantile of the potential outcome  $Y^d$  given  $X = x$ ,  $p(W) = p(w)$ , and  $S = 1$ . By Assumption 2,

$$F_{Y | D, X, p(W), S=1}(y | d, x, p(w), 1) = \int I\{\varphi(d, x, u) \leq y\} dF_{u | D=d, X=x, p(W)=p(w), S=1}$$

$$\begin{aligned}
&= \int I\{\varphi(d, x, u) \leq y\} dF_{u|X=x, p(W)=p(w), S=1} \\
&= Q_{Y^d|S=1}^{\tau^{-1}}(x, p(w)).
\end{aligned}$$

The unconditional quantile of the potential outcome is identified as the inverse of the integration over the marginal distributions of  $X$  and  $p(W)$  given  $S = 1$ :

$$\int \int Q_{Y^d|S=1}^{\tau^{-1}}(x, p(w)) dF_{x|(p(W)=p(w), S=1)} dF_{p(w)|S=1} = Q_{Y^d|S=1}^{\tau^{-1}}. \quad (4)$$

The difference between the quantiles under treatment and non-treatment yields the QTE:

$$\Delta_{S=1}^{\tau} = Q_{Y^1|S=1}^{\tau} - Q_{Y^0|S=1}^{\tau}. \quad (5)$$

Identification of  $\Delta_{S=1}^{\tau}$  hinges on the common support of the treatment in  $X$  and  $p(W)$  in the subpopulation with observed outcomes. We therefore impose a further assumption.

**Assumption 3.** *Common support in the treatment propensity score among the selected.*

(3a)  $c < \Pr(D = 1 | X = x, p(W) = p(w), S = 1) < 1 - c$ ,  $\forall x, p(w)$  in the support of  $X, p(W)$ , respectively, and  $c > 0$  (common support of  $D$  in  $X$  and  $p(W)$ ).

Assumption 3 states that the treatment propensity score conditional on being observed is bounded away from zero and one. It is obvious that Assumption (2b) is a necessary condition for Assumption 3 to hold. For example, if the outcomes of individuals with  $D = 1$  were never observed,  $\Pr(D = 1 | X, p(W), S = 1)$  would always be zero. Assumption (2b) is, however, not sufficient for (3). Consider the case that all individuals receiving treatment  $D = 1$  and having characteristics  $X = x$  are selected, which is not ruled out by (2b). That is,  $D = 1, X = x$  implies  $p(W) = 1$ , independent of  $Z$ . If  $p(W) < 1$  for  $D = 0$  and any other value of  $Z$  given  $X = x$ , it follows that  $\Pr(D = 1 | X = x, p(W) = 1) = 1$ . Thus,  $p(W) = 1$  perfectly predicts that  $D = 1$  conditional on  $X = x$  in the subpopulation with observed outcomes such that the common support assumption fails. Unless the selection probability conditional on  $D = 1, X = x$  is not smaller than one, identification requires that there exists some combination of  $(D = 0, Z = z)$  with  $p(W) = 1$  given  $X = x$ .

At this point, let us consider the special case that Assumption 3 is satisfied and  $p(W) = 1$  for some triples  $(D, X, Z)$ . Obviously, selection bias

is not an issue for these observations as  $E[Y | D = d, X = x, p(W) = 1, S = 1] = E[Y | D = d, X = x, p(W) = 1]$ . This allows identifying local treatment effects for the subpopulation with  $p(W) = 1$ , given that there is variation in the treatment state. It remains a priori unclear why this particular population should be of any policy interest. However, if one is willing to impose the strong restriction of treatment effect homogeneity across selection probabilities, i.e.,  $\Delta_{S=1}(x, p(w)) = \Delta_{S=1}(x) \forall p(w)$  in the support of  $p(W)$ , treatment effects can be identified for other populations as well conditional on common support in  $X$ . Identification based on  $p(W) = 1$  is known as “identification at infinity” and was discussed by Heckman (1990) and Andrews and Schafgans (1998). However, in empirical applications, observation with selection probabilities close to one might be rare and effect homogeneity in  $p(W)$  is a strong assumption that might not hold in reality. We therefore concentrate on a more general identification strategy using the whole distribution of  $p(W)$ .

After having established the identifying assumptions, we will now propose expressions for  $\Delta_{S=1}, \Delta_{S=1}^{\tau}$  based on IPW which can be used to build sample analogues required for estimation. Let  $\pi(X, p(W))$  denote the treatment propensity score, i.e., the probability of being treated conditional on  $X$  and  $p(W)$ ,  $\pi(X, p(W)) \equiv \Pr(D = 1 | X, p(W))$ . To control for selection into treatment, we will henceforth condition on  $\pi(X, p(W))$  instead of  $X$  and  $p(W)$ . Rosenbaum and Rubin (1983) have shown that conditioning on the treatment propensity score is equivalent to conditioning on the covariates directly, as both are balancing scores in the sense that they adjust the distributions of covariates in the groups of treated and controls. However, conditioning on  $\pi(X, p(W))$  will have the advantage that practical problems related to the nonparametric estimation based on high dimensional covariates, e.g., empty cells for particular combinations of covariate values, can be circumvented.

**Proposition 1** (Identification of Mean Effects on the Selected Subpopulation).

*Under Assumptions 1, 2, and 3, the ATE in the subpopulation with observed outcomes is identified by*

$$\Delta_{S=1} = E \left[ \frac{D \cdot Y}{\pi(X, p(W))} \middle| S = 1 \right] - E \left[ \frac{(1 - D) \cdot Y}{1 - \pi(X, p(W))} \middle| S = 1 \right]. \quad (6)$$

**Proof.** See Appendix A.1.

The ATE on the selected subpopulation is identified by reweighing the observed outcomes by the inverse of the conditional treatment probability given  $X$  and  $p(W)$ . An analogous approach identifies the quantiles and the QTE.

**Proposition 2** (Identification of quantiles in the selected subpopulation).

Under Assumptions 1, 2, and 3,  $Q_{Y^1|S=1}^\tau$  is an implicit function of

$$E \left[ \frac{D}{\pi(X, p(W))} \cdot I\{Y \leq Q_{Y^1|S=1}^\tau\} \middle| S = 1 \right] = F_{Y^1|S=1}(Q_{Y^1|S=1}^\tau) = \tau. \quad (7)$$

*Proof.* See Appendix A.2.

It follows that

$$Q_{Y^1|S=1}^\tau = \arg \text{zero}_y E \left[ \frac{D}{\pi(X, p(W))} \cdot (I\{Y < y\} - \tau) \middle| S = 1 \right],$$

which is a first order condition to

$$Q_{Y^1|S=1}^\tau = \arg \min_y E \left[ \frac{D}{\pi(X, p(W))} \cdot \rho_\tau(Y - y) \middle| S = 1 \right]. \quad (8)$$

$\rho_\tau(a) \equiv a \cdot (\tau - I\{a < 0\})$  denotes the check function, an asymmetric loss function suggested by Koenker and Bassett (1978) for quantile regression. An equivalent identification result holds for  $Q_{Y^0|S=1}^\tau$ , and it follows that  $\Delta_{S=1}^\tau = Q_{Y^1|S=1}^\tau - Q_{Y^0|S=1}^\tau$ . Based on reweighing observed outcomes by the inverse of the nested propensity score, we identify the ATE and QTEs in the selected subpopulation.

As noted by Newey (2007), without further assumptions, effects cannot be identified for other groups than the selected subpopulation, as  $Y$  is not even observed when  $S = 0$ . However, under particular common support conditions and conditional homoscedasticity of  $Y$ , the IPW framework even allows identifying the ATE on the total population ( $\Delta = E[Y^1] - E[Y^0]$ ), i.e., irrespective of selection. To this end, we make the following two assumptions.

**Assumption 4.** *Common support in the sample selection and treatment propensity scores.*

(4a)  $\Pr(S = 1 | D = d, X = x, Z = z) > c$ ,  $\forall x, z$  in the support of  $X, Z$ , respectively, and  $c > 0$  (positive sample selection propensity score).

(4b)  $c < \Pr(D = 1 | X = x, p(W) = p(w)) < 1 - c$ ,  $\forall x, p(w)$  in the support of  $X, p(W)$ , respectively, and  $c > 0$  (common support in the treatment propensity score).

(4a) states that the sample selection propensity score is bounded away from zero, which is stronger than (2b). Effects on the total population could not be identified if there existed individuals with a sample selection

propensity score equal to zero as this would rule out suitable comparisons in the subpopulation with observed outcomes. (4b) states that there must be common support in the treatment propensity score in the population.

**Assumption 5.** *Separability of observed and unobserved terms.*

(5a)  $Y = \varphi(D, X) + U$  (separability).

Assumption 5 ensures homoscedasticity of  $Y$  given  $(D, X)$ , which is required for the subsequent proposition. Nonparametric sample selection models with additive unobserved terms have also been considered in Das et al. (2003). Note that while nonseparable models allow for effect heterogeneity w.r.t. unobserved terms even conditional on  $X$ , models with separability do not. That is, Assumption 5 comes with the cost of a decreased generality of the model. The plausibility of this restriction has to be judged in the empirical application at hand. In particular, it has to be justified that the observed covariates are sufficiently rich such that the treatment effect is homogenous given this information. This may be the case when assessing a new medical treatment where all relevant socioeconomic and health-related characteristics of the patients are measured prior to the intervention. An example where the assumption is less likely to hold is the evaluation of the returns to schooling or training, where we would suspect the effectiveness of the intervention to vary with unobserved ability.

**Proposition 3** (Identification of mean effects on the total population).

*Under Assumptions 1, 2, 4, and 5, the ATE on the total population is identified by*

$$\Delta = E \left[ \frac{S \cdot D \cdot Y}{p(W) \cdot \pi(X, p(W))} \right] - E \left[ \frac{S \cdot (1 - D) \cdot Y}{p(W) \cdot (1 - \pi(X, p(W)))} \right]. \quad (9)$$

**Proof.** See Appendix A.3.

The ATE on the total population is identified based on reweighing observations (additionally to the inverse treatment propensity score) by the inverse of the sample selection propensity score, i.e., by using the relative likelihood of a particular triple  $(D, X, Z)$  to appear in the total population as weighting function. It may seem surprising that identification is possible even though outcomes are only partially observed and the observed outcomes do generally not allow inferring on the unobserved outcomes. That is,  $E[Y | D = d, X = x, p(W) = p(w), S = 1] \neq E[Y | D = d, X = x, p(W) = p(w), S = 0]$  due to different conditional distributions of the unobserved term  $U$ . However, Assumptions (2c) and (5) imply that

$\Delta_{S=1}(x, p(w)) = \Delta_{S=0}(x, p(w))$ . To see this, note that by Assumption (2c),  $F_{U|D=d, X=x, p(W)=p(w), S=s} = F_{U|X=x, p(W)=p(w), S=s}$  for  $s \in \{0, 1\}$  such that

$$\begin{aligned}\Delta_{S=1}(x, p(w)) &= \int [\varphi(1, x) + u] dF_{u|X=x, p(W)=p(w), S=1} \\ &\quad - \int [\varphi(0, x) + u] dF_{u|X=x, p(W)=p(w), S=1}, \\ \Delta_{S=0}(x, p(w)) &= \int [\varphi(1, x) + u] dF_{u|X=x, p(W)=p(w), S=0} \\ &\quad - \int [\varphi(0, x) + u] dF_{u|X=x, p(W)=p(w), S=0}.\end{aligned}$$

$\Delta_{S=1}(x, p(w))$  and  $\Delta_{S=0}(x, p(w))$  only differ with respect to the integrals over different conditional distributions of  $U$  given  $S = 1$  and  $S = 0$ , which cancel out in the subtractions due to the additivity assumption. Thus,  $\Delta_{S=1}(x, p(w)) = \Delta_{S=0}(x, p(w))$ . Therefore, reweighing the conditional treatment effects in the subpopulation with observed outcomes according to the distribution of  $(D, X, Z)$  in the total population identifies  $\Delta$ .

It seems useful to confront our results to Wooldridge (2002, 2007) who discusses IPW M-estimation of missing data models. Wooldridge considers the estimation of the general objective function  $m(A; \theta)$ , where  $A$  denotes a data matrix and  $\theta$  is the parameter of interest. The latter is identified by the moment condition  $E\left[\frac{S}{p(W)} m(A; \theta)\right] = 0$ . By defining  $m(A; \theta)$  as  $\left(\frac{D}{\pi(X, p(W))} - \frac{(1-D)}{1-\pi(X, p(W))}\right) \cdot (Y - \theta)$ , it follows that  $E\left[\frac{S}{p(W)} \cdot \left(\frac{D}{\pi(X, p(W))} - \frac{(1-D)}{1-\pi(X, p(W))}\right) \cdot (Y - \theta)\right] = 0$  such that  $\theta$  identifies the ATE on the total population. At a first glance, our results appear to be a special case.

However, the framework of Wooldridge (2002, 2007) is somewhat different because it does not consider sample selection on unobservables such that the sample selection propensity score  $p(W)$  does not enter the objective function  $m(A; \theta)$ . That is,  $V$ , the unobserved term in  $S$  must not be related with  $U$ , the unobserved factor in  $Y$ , whereas instrument  $Z$  may be related with  $U$ . In the selection on unobservables framework treated in this paper (which also underlies the classic sample selection literature)  $Z$  must not be related with  $V$  and  $U$ , but  $V$  may be related with  $U$ , see Fitzgerald et al. (1998) for a discussion of these distinct assumptions. For the same reason, our sample selection problem also differs from Robins and Rotnitzky (1995), Robins et al. (1995), and Rotnitzky and Robins (1995), who consider IPW adjusted regression under selection on observables.

Furthermore, we can link our work to identification based on IPW under the CIA, see for instance Hirano et al. (2003) and Firpo (2007).

The validity of the CIA in the absence of sample selection implies that the treatment effect is unconfounded conditional on the treatment propensity score with respect to  $X$  alone. In our framework, we need to condition on both  $X$  and  $p(W)$  to control for selection into the subpopulation with observed outcomes *and* into the treatment.

## 2.4. Further Target Populations

We have discussed the identification of treatment effects on the subpopulation with observed outcomes and on the total population. However, depending on the evaluation problem, different target populations might be relevant from a policy perspective. For example, Lee (2009) and Zhang et al. (2008) focus on the subpopulation of those being selected irrespective of the treatment assignment. Let  $S^d$  denote the potential sample selection indicator for treatment  $D = d$ . If one is willing to assume that the sample selection increases uniformly in the treatment (see for instance Lee, 2009 and Lechner and Melly, 2007), i.e.,  $\Pr(S^1 \geq S^0) = 1$ , then those observations with  $(S = 1, D = 0)$  are always selected irrespective of the treatment assignment, satisfying  $(S^1 = 1, S^0 = 1)$ . The always selected, or “always takers” in the notation of Imbens and Angrist (1994), are the *nontreated* individuals in the subpopulation with observed outcomes.

Hirano et al. (2003) discuss the identification of weighted ATEs based on IPW, which provides a general framework for the identification of treatment effects on different target populations. Translated to our sample selection framework their results imply that

$$\Delta_{g|S=1} = \frac{1}{E[g|S=1]} \cdot E \left[ \frac{D \cdot Y \cdot g}{\pi(X, p(W))} - \frac{(1-D) \cdot Y \cdot g}{1 - \pi(X, p(W))} \middle| S=1 \right],$$

where  $g$  is a general weighting function. For the always selected, the weight to be used is the propensity not to receive the treatment,  $1 - \pi(X, p(W))$ , because reweighing the conditional effect  $\Delta_{S=1}(x, p(w))$  and integrating over the distributions of  $X$  and  $p(W)$  in the selected sample yields the ATE on the always selected, denoted as  $\Delta_{S=1, D=0}$ :

$$\begin{aligned} \Delta_{S=1, D=0} &= \int \int \Delta_{S=1}(x, p(w)) dF_{x|p(W)=p(w), D=0, S=1} dF_{p(w)|D=0, S=1} \\ &= \int \int \Delta_{S=1}(x, p(w)) (1 - \pi(x, p(w))) dF_{x|p(W)=p(w), S=1} dF_{p(w)|S=1} / \\ &\quad \int \int (1 - \pi(x, p(w))) dF_{x|p(W)=p(w), S=1} dF_{p(w)|S=1}. \end{aligned}$$



Therefore,  $\Delta_{S=1,D=0}$  is identified by

$$\Delta_{S=1,D=0} = \frac{1}{\Pr(D=0 \mid S=1)} \cdot E \left[ D \cdot Y \cdot \frac{1 - \pi(X, p(W))}{\pi(X, p(W))} - (1 - D) \cdot Y \mid S=1 \right],$$

where  $\Pr(D=0 \mid S=1) = E[1 - \pi(X, p(W)) \mid S=1]$ . All observations ( $S=1, D=1$ ) are reweighed by  $\frac{1 - \pi(X, p(W))}{\pi(X, p(W))}$  such that they are comparable to the always selected ( $S=1, D=0$ ) in terms of the treatment propensity score. Similarly, the quantile  $Q_{Y^1 \mid S=1, D=0}^\tau$  is an implicit function of

$$E \left[ \frac{D}{\Pr(D=0 \mid S=1)} \cdot \frac{1 - \pi(X, p(W))}{\pi(X, p(W))} \cdot I\{Y \leq Q_{Y^1 \mid S=1, D=0}^\tau\} \mid S=1 \right],$$

see also the discussion on the identification of quantile treatment effects on the treated (QTET) in Firpo (2007). An equivalent result holds for  $Q_{Y^0 \mid S=1, D=0}^\tau$ , which implies the identification of  $\Delta_{S=1, D=0}^\tau$ . Note that Assumption 3 can be relaxed to  $c < \Pr(D=1 \mid X, p(W), S=1)$ ,  $c > 0$ , which suffices for the exclusion of arbitrarily large weights  $\frac{1 - \pi(X, p(W))}{\pi(X, p(W))}$ .

By the same logic, the ATE on those with ( $S=1, D=1$ ) is identified by weighting with  $\pi(X, p(W))$ . Given that uniformity of  $S$  in  $D$  holds, this group is made up by two subpopulations, namely the always selected ( $S^1=1, S^0=1$ ) and those individuals who are selected under treatment, but would not be under non-treatment ( $S^1=1, S^0=0$ ). In the spirit of Imbens and Angrist (1994), we refer to this latter group as compliers, where compliance means that the selection state reacts on the treatment assignment. For example, when evaluating the returns to a training, the compliers are those who switch into employment when being placed into a training. Evaluating the effects on the potential wages of individuals who change their labor market behavior in light of the treatment may be of great policy relevance and compliers appear to be an interesting population in many other problems, too. We can identify the ATE on the compliers, denoted as  $\Delta_{S^1=1, S^0=0}$ , by making the following observation:

$$\Delta_{S=1} = \Delta_{S=1, D=1} \cdot \Pr(D=1 \mid S=1) + \Delta_{S=1, D=0} \cdot \Pr(D=0 \mid S=1),$$

where

$$\begin{aligned} \Delta_{S=1, D=1} &= \Delta_{S^1=1, S^0=1} \cdot \Pr(S^1=1, S^0=1 \mid S=1, D=1) \\ &\quad + \Delta_{S^1=1, S^0=0} \cdot (1 - \Pr(S^1=1, S^0=1 \mid S=1, D=1)) \end{aligned}$$

$$\begin{aligned}
&= \Delta_{S=1, D=0} \cdot \frac{\Pr(S=1 \mid D=0)}{\Pr(S=1 \mid D=1)} \\
&\quad + \Delta_{S^1=1, S^0=0} \cdot \left(1 - \frac{\Pr(S=1 \mid D=0)}{\Pr(S=1 \mid D=1)}\right).
\end{aligned}$$

The first and second equalities follow from the law of total probability. The third equality holds because of  $\Pr(S^1 \geq S^0) = 1$  such that the always selected are one subpopulation in  $(S=1, D=1)$ . Their fraction is  $\frac{\Pr(S=1 \mid D=0)}{\Pr(S=1 \mid D=1)}$ , i.e., the share of individuals that would even be selected without treatment among those selected under the treatment. Therefore, the remaining fraction  $1 - \frac{\Pr(S=1 \mid D=0)}{\Pr(S=1 \mid D=1)}$  must be made up by compliers, see also Lee (2009). This allows identifying the ATE on the compliers by

$$\begin{aligned}
\Delta_{S^1=1, S^0=0} &= \Delta_{S=1, D=1} \cdot \left(1 - \frac{\Pr(S=1 \mid D=0)}{\Pr(S=1 \mid D=1)}\right)^{-1} \\
&\quad - \Delta_{S=1, D=0} \cdot \frac{\Pr(S=1 \mid D=0)}{\Pr(S=1 \mid D=1)} \cdot \left(1 - \frac{\Pr(S=1 \mid D=0)}{\Pr(S=1 \mid D=1)}\right)^{-1}.
\end{aligned}$$

The framework of weighted treatment effects could be used to identify the effects on further target populations, but this is beyond the scope of this paper. The empirical application will focus on the subpopulation with observed outcomes.

### 3. ESTIMATION

In this section, we briefly discuss the estimation of treatment effects based on our identification results. Note that the selection and treatment propensity scores are unknown and have to be estimated in order to be used in the weighting functions of the estimators of the ATEs and QTEs which we denote by  $\hat{\Delta}_{S=1}, \hat{\Delta}_{S=1}^{\tau}, \hat{\Delta}$ . Furthermore, let  $\hat{p}(W), \hat{\pi}(X, \hat{p}(W))$  denote the estimates of the sample selection propensity score  $p(W)$  and the treatment propensity score  $\pi(X, p(W))$ , respectively. A general 3-step estimation approach can be outlined as follows:

- Estimate  $\hat{p}(W)$  by regressing  $S$  on  $D, X, Z$ ;
- Estimate  $\hat{\pi}(X, \hat{p}(W))$  by regressing  $D$  on  $X$  and  $\hat{p}(W)$ ;
- Estimate  $\hat{\Delta}_{S=1}, \hat{\Delta}_{S=1}^{\tau}, \hat{\Delta}$  by the normalized sample analogues of (6), (8), and (9).

Concerning the estimation of the propensity scores in steps (a) and (b), Ahn and Powell (1993) and Hirano et al. (2003) have proposed nonparametric methods based on kernel regression and series

approximation, respectively. However, the empirical literature mainly uses parametric specifications based on logit or probit models, which do not require the choice of smoothing parameters, but can be made arbitrarily flexible by including interaction and higher order terms. We also follow this strategy in the empirical application presented in Section 5.

For the estimation of the treatment effects in (c), we suggest to use the normalized sample analogues of the identification results; for example, the normalized estimator of the ATE on the selected is

$$\hat{\Delta}_{S=1} = \sum_{i|S=1}^n \frac{D_i \cdot Y_i}{\hat{\pi}(X_i, \hat{p}(W_i))} \bigg/ \sum_{j|S=1}^n \frac{D_j}{\hat{\pi}(X_j, \hat{p}(W_j))} - \sum_{i|S=1}^n \frac{(1 - D_i) \cdot Y_i}{1 - \hat{\pi}(X_i, \hat{p}(W_i))} \bigg/ \sum_{j|S=1}^n \frac{(1 - D_j)}{1 - \hat{\pi}(X_j, \hat{p}(W_j))}.$$

Here, the normalizations  $\sum_{j|S=1}^n \frac{D_j}{\hat{\pi}(X_j, \hat{p}(W_j))}$  and  $\sum_{j|S=1}^n \frac{(1-D_j)}{1-\hat{\pi}(X_j, \hat{p}(W_j))}$  guarantee that the weights add up to unity. This may entail better finite sample properties of the estimator, see for instance the discussion in Imbens (2004) and Busso et al. (2009). Based on the results of Hirano et al. (2003), Firpo (2007), and Hahn and Ridder (2013), Appendix A.4 provides the asymptotic variance of the 3-step IPW estimators  $\hat{\Delta}_{S=1}$ ,  $\hat{\Delta}_{S=1}^{\tau}$  for the general case that the propensity scores are estimated nonparametrically.

Besides the asymptotic results, the bootstrap may also be used as inference method for IPW, see Hirano et al. (2003) and Firpo (2007). In the application, we therefore repeatedly draw bootstrap samples of size  $n$  with replacement out of the original data in order to estimate the distribution of the treatment effect estimates based on which the standard errors are constructed. In each sample, all estimation steps are conducted such that the uncertainty coming from both effect and propensity score estimation is accounted for. As a final remark on estimation, it is worth noting that PSM on the nested score may be used as an alternative method to IPW. Both classes of estimators rely on the same identifying assumptions, see Lechner (2007), and should therefore give similar results in large samples. For this reason, we also consider PSM in the application as well as the simulations.

#### 4. EMPIRICAL APPLICATION

In this section we estimate the effect of high school graduation on female wages using a subsample of the U.S. CPS data of Mulligan and Rubinstein (2008). In contrast to the original data, our sample only contains females graduating from high school or dropping out after 9 to 11 years of schooling. It consists of a repeated cross-section that covers

the years 1975 to 1979 and contains information on white females aged between 25 and 54. Individuals are classified as working ( $S = 1$ ) if they work 35+ hours per week and at least 50 weeks during the year. Self-employed and persons in the military, agriculture, or private household sectors as well as individuals with inconsistent reports on earnings or with allocated earnings are excluded from the sample with observed wages, see Mulligan and Rubinstein (2008) for further details. The outcome variable ( $Y$ ) is female's hourly wage, which is computed based on total annual earnings which are deflated by the U.S. Consumer Price Index (CPI).

We are interested in the ATE and QTEs of graduating from high school ( $D = 1$ ) vs. receiving 9 to 11 years of schooling without high school graduation ( $D = 0$ ) on the wages of working females. The evaluation sample consists of 67,848 observations, thereof 52,354 high school graduates and 15,494 high school drop-outs; 16,550 graduates and 3,598 drop-outs are observed to work according to the definition of Mulligan and Rubinstein (2008). In addition to education, the data include information on potential work experience, the marital status, and regional dummies, which serve as covariate vector  $X$ . Finally, the number of children aged 0–6 and its interactions with the marital status are used as exclusion restrictions  $Z$ .

Table 2 reports descriptive statistics, namely, the means and standard deviations of the key variables, for the evaluation sample and subsamples defined upon selection as well as selection and treatment, respectively. The descriptives indicate that sample selection is non-random. Notably, the potential experience among the selected is half-a-year higher than in the entire evaluation sample. Furthermore, the former are 13 percentage

**TABLE 2** Descriptive statistics

Variable	Entire sample		Selected ( $S = 1$ )		$S = 1, D = 1$		$S = 1, D = 0$	
	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.	Mean	St. Dev.
Potential experience	6.508	8.980	7.071	8.899	6.491	8.894	9.741	8.420
Married (binary)	0.830	0.376	0.702	0.457	0.701	0.458	0.707	0.455
Separated (binary)	0.022	0.147	0.027	0.162	0.026	0.158	0.034	0.180
Widowed (binary)	0.030	0.172	0.043	0.202	0.039	0.194	0.058	0.233
Divorced (binary)	0.080	0.272	0.145	0.352	0.141	0.348	0.165	0.371
Never married (binary)	0.038	0.190	0.083	0.276	0.093	0.291	0.037	0.189
Midwest (binary)	0.291	0.454	0.293	0.455	0.297	0.457	0.278	0.448
South (binary)	0.278	0.448	0.307	0.461	0.298	0.457	0.350	0.477
West (binary)	0.193	0.395	0.185	0.388	0.188	0.391	0.171	0.377
# children < 6 yrs.	0.338	0.670	0.151	0.430	0.155	0.434	0.136	0.403
High school grad ( $D$ )	0.772	0.420	0.821	0.383	1.000	0.000	0.000	0.000
Working ( $S$ )	0.297	0.457	1.000	0.000	1.000	0.000	1.000	0.000
Hourly wage ( $Y$ )	—		11.557	4.916	11.797	5.119	9.957	4.334
Number of obs.	67,848		20,148		16,550		3,598	

points less likely to be married and 7 percentage points more likely to be divorced. This is intuitive, as non-married females cannot rely on a spouse as alternative source of income. Also the lower average number of children under 6 among working females appears plausible. The table also shows that in particular the potential experience varies importantly across treated and nontreated observations with observed outcomes such that the treatment choice appears selective, too.

In our estimation, we use flexible probit specifications for the propensity scores that include higher order and interaction terms.  $p(W)$  is a function of the treatment dummy (high school graduation), marital status and the number of children aged 0–6 along with interactions, number of kids aged 0–6 squared, a potential work experience cubic interacted with education dummies, and the regional dummies.  $\pi(X, p(W))$  is a function of  $p(W)$ , the potential work experience cubic, marital status, and the region. In the sample selection equation, the coefficients on high school graduation, the marital status, the number of children, the region, and interaction terms between experience and high school graduation are significant at the 5% level, and in the treatment equation, all coefficients, including the one on the predicted sample selection propensity score, are significant at the 1% level.

The common support or overlap in the treatment propensity score distributions of treated and nontreated units is of crucial importance. An insufficient overlap would point to a lack of appropriate comparisons across treatment groups. The histograms of  $\hat{\pi}(X_i, \hat{p}(W_i))$  for  $D = 1$  and  $D = 0$  presented in Fig. 1 reveal that the common support is quite satisfactory. In fact, both treatment groups contain observations over the entire theoretical support of the propensity score.

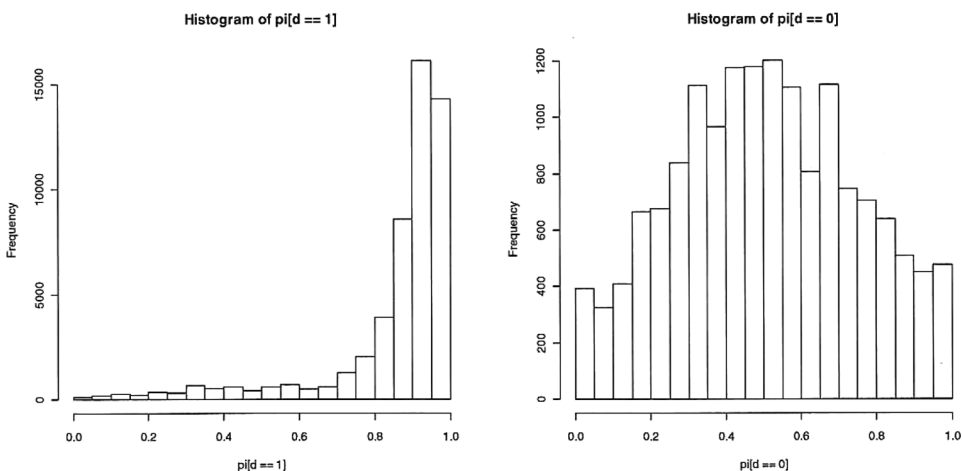


FIGURE 1 Estimated treatment propensity scores for  $D = 1$  and  $D = 0$ .

As some propensity score estimates are close to the boundaries, we trim these values to a maximum of 0.99 and a minimum of 0.01 to avoid arbitrarily large weights which might entail instability of the IPW estimator, see Khan and Tamer (2010). Furthermore, wage outliers are trimmed when estimating the ATE and left unchanged when estimating QTEs in the same manner as in Mulligan and Rubinstein (2008).

In addition to IPW, we estimate the ATE on the observed subpopulation using PSM. To be specific, we use two-nearest-neighbor caliper matching (see Sekhon, 2011) where the caliper defines the maximally acceptable distance in any match's propensity score. This procedure eliminates those matches that are not comparable in terms of their treatment probabilities, i.e., lie outside the support. We set the caliper to 0.25 standard deviations of the estimated treatment propensity score, but due to the decent common support, no observations have to be dropped. After-matching balance tests indicate that balance is considerably increased, suggesting that treated and nontreated matches are comparable with respect to the distribution of the covariates and the estimated sample selection propensity score. We use 999 bootstrap replications to compute standard errors and  $p$ -values of the IPW estimators. PSM standard errors are estimated by the (within treatment group) matching-based variance estimator suggested by Abadie and Imbens (2006), which, however, does not account for uncertainty in the estimation of the propensity scores.

Table 3 provides the ATE estimates ( $\hat{\Delta}_{S=1}$ ) and standard errors (s.e.) of the semiparametric IPW and PSM procedures. The highly significant effects suggest that graduating from high school increases the average hourly wage by 1.93 to 1.98 USD or 17.1%. The estimate of the parametric two step heckit procedure (see Heckman, 1976), which is also provided in the table, is somewhat lower (1.63 USD or 14.1%). The results are in line with the partial effect of high school graduation vs. non-graduation reported in Table A.2 of Mulligan and Rubinstein (2008), which is 16.7% based on the complete CPS sample with all education categories.

**TABLE 3** ATE estimates (increase of hourly wage in USD)

	IPW	PSM	Heckit		Worst case bounds	Bounds w. assumptions
$\hat{\Delta}_{S=1}$	1.980	1.934	1.626	Identified set	[−29.040, 13.026]	[1.440, 2.061]
(s.e.)	0.173	0.450	0.115	(s.e.'s of bounds)	(1.815, 0.395)	(0.158, 0.695)
$p$ -value	0.000	0.000	0.000	Confidence regions	[−32.597, 13.800]	[1.130, 3.423]

Note: Standard errors (s.e.'s) of IPW and worst case bounds are based on 999 bootstrap replications.

S.e.'s of bounds with assumptions are based on 999 subsampling draws.

S.e.'s of PSM are computed using the Abadie and Imbens (2006) variance estimator.

S.e.'s of heckit is based on asymptotic theory.

To assess the credibility of our results, we compare them to the worst case bounds (see Manski, 1989 and 1994, for an introduction to partial identification) on  $\Delta_{S=1}$  when neither controlling for sample selection nor treatment selection. Note that

$$\begin{aligned}
 \Delta_{S=1}^{UB} &= E[Y | D = 1, S = 1] \cdot \Pr(D = 1 | S = 1) + UB \cdot (1 - \Pr(D = 1 | S = 1)) \\
 &\quad - LB \cdot \Pr(D = 1 | S = 1) \\
 &\quad - E[Y | D = 0, S = 1] \cdot (1 - \Pr(D = 1 | S = 1)) \\
 &\geq \Delta_{S=1} \geq [Y | D = 1, S = 1] \cdot \Pr(D = 1 | S = 1) \\
 &\quad + LB \cdot (1 - \Pr(D = 1 | S = 1)) \\
 &\quad - UB \cdot \Pr(D = 1 | S = 1) - E[Y | D = 0, S = 1] \\
 &\quad \cdot (1 - \Pr(D = 1 | S = 1)) = \Delta_{S=1}^{LB},
 \end{aligned}$$

where  $\Delta_{S=1}^{UB}$  and  $\Delta_{S=1}^{LB}$  denote the upper and lower bound of the ATE.  $UB$  and  $LB$  are the upper and lower bound of hourly wages, which are set to the maximum and minimum observed wages in the data,  $\max(Y | S = 1)$ ,  $\min(Y | S = 1)$ , respectively. For estimation, we simply take the sample analogues of  $\Pr(D = 1 | S = 1)$  and  $E[Y | D = d]$ . Not surprisingly, the estimated bounds are quite uninformative as the admissible ATEs range from  $-29.040$  to  $13.026$ . We bootstrap the lower and upper bound 999 times in order to estimate their standard errors and compute the confidence interval  $[\hat{\Delta}_{S=1}^{LB} - 1.96 \cdot \hat{\sigma}_{LB}, \hat{\Delta}_{S=1}^{UB} + 1.96 \cdot \hat{\sigma}_{UB}]$ , where  $\hat{\Delta}_{S=1}^{LB}$ ,  $\hat{\Delta}_{S=1}^{UB}$ ,  $\hat{\sigma}_{LB}$ ,  $\hat{\sigma}_{UB}$  are the estimates of the ATE bounds and their respective standard errors. This confidence interval covers the true ATE on the working with at least 0.95 probability.

To tighten the bounds, we assume the CIA to hold conditional on  $\pi(X) = \Pr(D = 1 | X)$ , which is implied by Assumption 1, and impose the uniformity assumption of Lechner and Melly (2007) and Lee (2009); that is,  $\Pr(S^1 \geq S^0) = 1$  (see also the last section), implying that everyone working without graduation would also work with graduation. Lee bounds the treatment effect for the always selected with  $S^0 = 1$ ,  $S^1 = 1$ , which are those who work irrespective of education. Under the CIA and the uniformity assumption  $E[Y | D = 0, \pi(X)]$  is equal to the expected potential outcome for the always selected under non-graduation,  $E[Y^0 | \pi(X), S^0 = 1, S^1 = 1]$ .  $E[Y | D = 1, \pi(X)]$  is a weighted average of outcomes of always selected and compliers, i.e., individuals who work with graduation but would not without graduation ( $S^0 = 0, S^1 = 1$ ). That is,

$$\begin{aligned}
 E[Y | D = 1, \pi(X)] \\
 &= E[Y | D = 1, \pi(X), S^0 = 1, S^1 = 1] \cdot (1 - c)
 \end{aligned}$$

$$\begin{aligned}
& + E[Y | D = 1, \pi(X), S^0 = 0, S^1 = 1] \cdot c, \\
& = E[Y^1 | \pi(X), S^0 = 1, S^1 = 1] \cdot (1 - c) + E[Y^1 | \pi(X), S^0 = 0, S^1 = 1] \cdot c,
\end{aligned}$$

where  $c$  denotes the probability to be a complier given the propensity score,  $Pr(S^0 = 0, S^1 = 1 | \pi(X))$ .

Thus, the expected potential outcome for the always selected under graduation,  $E[Y^1 | \pi(X), S^0 = 0, S^1 = 1]$ , can be bounded by taking the expectation of the upper or lower share of  $Y | D = 1, S = 1, \pi(X)$  that corresponds to the probability to be an always selected, which is  $1 - c = 1 - \frac{Pr(S=1 | D=1, \pi(X)) - Pr(S=1 | D=0, \pi(X))}{Pr(S=1 | D=1, \pi(X))}$ , see Lee (2009) for further details. The upper and lower bounds on the ATE for the always selected,  $\Delta_a$ , are identified by

$$\begin{aligned}
\Delta_a^{UB} &= \int \{E[Y | D = 1, S = 1, \pi(X) = \pi(x), Y \geq Q_Y^c] \\
&\quad - E[Y | D = 0, S = 1, \pi(X) = \pi(x)]\} dF_{\pi(X) | S=1}, \\
\Delta_a^{LB} &= \int \{E[Y | D = 1, S = 1, \pi(X) = \pi(x), Y \leq Q_Y^{1-c}] \\
&\quad - E[Y | D = 0, S = 1, \pi(X) = \pi(x)]\} dF_{\pi(X) | S=1},
\end{aligned}$$

where  $Q_Y^\tau$  denotes the  $\tau$ th quantile of  $Y$ .

As we want to estimate the bounds for the entire population with observed outcomes ( $S = 1$ ), we also need to bound the counterfactual of  $E[Y | D = 1, S = 1, \pi(X)]$ , which is

$$\begin{aligned}
E[Y^0 | D = 1, S = 1, \pi(X)] &= (1 - c) \cdot E[Y^0 | \pi(X), S^0 = 1, S^1 = 1] \\
&\quad + c \cdot E[Y^0 | \pi(X), S^0 = 0, S^1 = 1].
\end{aligned}$$

Due to the uniformity assumption the counterfactual for the always selected,  $E[Y^0 | \pi(X), S^0 = 1, S^1 = 1]$ , is  $E[Y | D = 0, \pi(X), S^0 = 1, S^1 = 1] = E[Y | D = 0, S = 1, \pi(X) = \pi(x)]$  and observed. However,  $E[Y^0 | \pi(X), S^0 = 0, S^1 = 1]$  is unknown as complier outcomes are not observed for  $D = 0$ . We define the upper bound of  $E[Y^0 | \pi(X), S^0 = 0, S^1 = 1]$  as  $E[Y | D = 0, S = 1, \pi(X) = \pi(x)]$ , assuming that observed compliers would on average not earn more without graduation than the always selected. The latter would be employed with and without graduation and are therefore likely to be more motivated and/or able than the compliers. Zhang et al. (2008) argue that ability tends to be positively correlated with wages, and thus, this assumption appears to be plausible. Also Lechner and Melly (2007) assume positive selection with respect to wages.

We define the lower bound of  $E[Y^0 | \pi(X), S^0 = 0, S^1 = 1]$  as the minimum wage that is observed among working,  $\min(Y | S = 1)$ . Then, the



upper and lower bounds of the ATE on the always selected and compliers,  $\Delta_{ac}$ , are identified by

$$\begin{aligned}\Delta_{ac}^{UB} &= \int \{E[Y | D = 1, S = 1, \pi(X) = \pi(x)] \\ &\quad - E[Y | D = 0, S = 1, \pi(X) = \pi(x)] \\ &\quad \cdot (1 - c) - \min(Y | S = 1) \cdot c\} dF_{\pi(X) | S=1}, \\ \Delta_{ac}^{LB} &= \int \{E[Y | D = 1, S = 1, \pi(X) = \pi(x)] \\ &\quad - E[Y | D = 0, S = 1, \pi(X) = \pi(x)]\} dF_{\pi(X) | S=1}.\end{aligned}$$

Finally,  $\Delta_{S=1}$  is partially identified by

$$\begin{aligned}\Delta_{ac}^{UB} \cdot \Pr(D = 1 | S = 1) + \Delta_a^{UB} \cdot (1 - \Pr(D = 1 | S = 1)) &\geq \Delta_{S=1} \\ \geq \Delta_{ac}^{LB} \cdot \Pr(D = 1 | S = 1) + \Delta_a^{LB} \cdot (1 - \Pr(D = 1 | S = 1)).\end{aligned}$$

We estimate  $\pi(X)$  using a probit model and denote the estimated propensity score as  $\hat{\pi}(X)$ . Also  $\Pr(S = 1 | D = d, \pi(X))$  is estimated by a probit regression of  $S | D = d$  on  $\hat{\pi}(X)$  and a constant.  $E[Y | D = 1, S = 1, \pi(X) = \pi(x), Y \geq Q_Y^c]$ , and  $E[Y | D = 1, S = 1, \pi(X) = \pi(x), Y \leq Q_Y^{1-c}]$  are estimated by averaging over the predictions of linear quantile regressions of  $Y | D = 1$  on the polynomial  $\sum_{p=0}^3 \hat{\pi}(X)^p$  and  $E[Y | D = d, S = 1, \pi(X) = \pi(x)]$  by averaging over the predictions of a linear mean regression of  $Y | D = d$  on  $\sum_{p=0}^3 \hat{\pi}(X)^p$ .  $\Delta_a^{UB}, \Delta_a^{LB}, \Delta_{ac}^{UB}, \Delta_{ac}^{LB}$  are estimated by matching on  $\hat{\pi}(X)$ . To compute the confidence intervals, we draw 999 subsamples without replacement, see Politis et al. (1999), of subsample size 20,000. Under the CIA and the uniformity assumption the identified set is quite informative and positive. The ATE's lower bound is significantly different from zero. Notably, the IPW and PSM point estimates lie within the estimated bounds and therefore do not contradict the results obtained from partial identification.

Finally, Table 4 reports the QTE estimates based on IPW for the 0.1th to 0.9th quantile of potential wages. The effects vary importantly across

**TABLE 4** QTE estimates (increase of hourly wage in USD)

$\tau$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\hat{\Delta}_{S=1}^{\tau}$	1.119	1.148	1.436	1.900	2.199	2.199	2.443	2.452	2.671
(s.e.)	0.225	0.226	0.213	0.212	0.231	0.218	0.225	0.216	0.229
$p$ -value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note: Standard errors (s.e.'s) of IPW are based on 999 bootstrap replications.

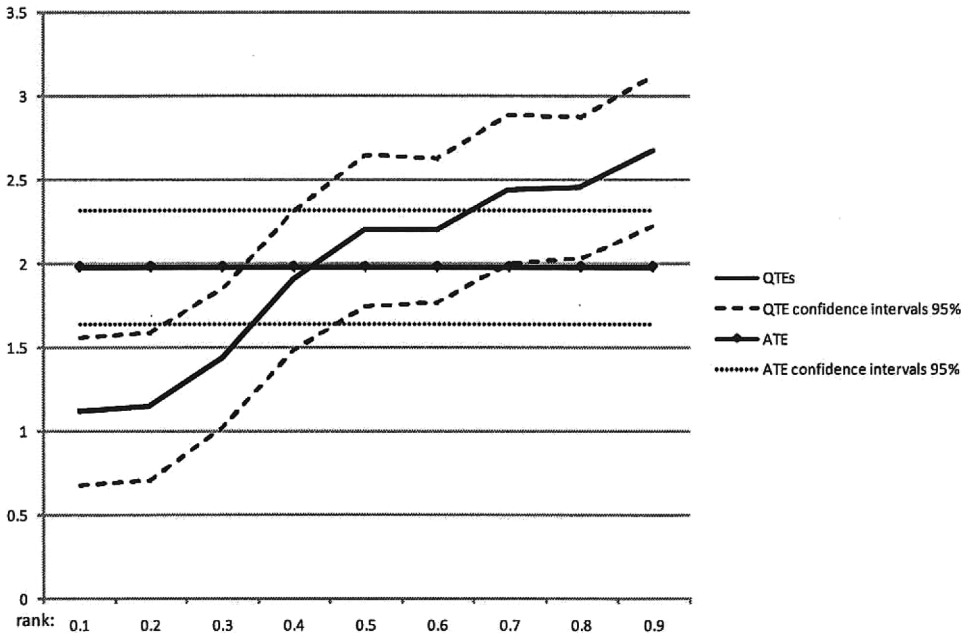


FIGURE 2 QTE and ATE estimates (in USD) and confidence intervals.

different parts of the wage distribution. The results suggest that those with comparably large hourly wages benefit most while those with little wages benefit least from a high school graduation, given that the rank stability assumption holds. Figure 2 displays the QTEs and ATE along with pointwise confidence intervals to show the variation in the effects across quantiles.

## 5. SIMULATIONS

This section presents Monte Carlo simulations based on linear and nonlinear sample selection models to examine the finite sample properties of IPW and PSM relative to parametric maximum likelihood and two-step procedures, and the naive estimator (i.e., the difference in the sample means of the observed treated and observed nontreated outcomes). The effect estimates refer to the subpopulation with observed outcomes.

The first data generating process (DGP) is a classic linear selection model with bivariate normally distributed errors the covariance of which is set to 0.8:

$$Y_i = \alpha_1 D_i + \alpha_2 X_i + U_i,$$

$$Y_i \text{ is observed if } S_i = 1,$$

$$S_i = I\{\beta_1 D_i + \beta_2 X_i + \beta_3 Z_i + V_i > 0\},$$

$$D_i = I\{\gamma_1 X_i + \varepsilon_i > 0\},$$

$$X, Z \sim N(0, 1), U, V, \varepsilon \sim N(0, 2), \quad \text{Cov}(U, V) = 0.8, \quad \text{Cov}(U, \varepsilon) = 0,$$

$$\alpha_1 = \alpha_2 = 1, \quad \beta_1 = \beta_2 = 0.25, \quad \beta_3 = \gamma_1 = 0.5.$$

We run 1,000 Monte Carlo replications and consider two sample sizes ( $n = 700, 2800$ ). The smaller sample size is similar to the prominent data set of female wages and labor supply in Mroz (1987) which has been re-analyzed several times in the empirical sample selection literature, see for instance Newey et al. (1990), Ahn and Powell (1993), and Greene (2003). The larger sample size is comparable to more recent studies such as Martins (2001). We estimate the median effect and the ATE by IPW and trim (probit-based) propensity score estimates that are larger than 0.95 (0.975) and smaller than 0.05 (0.025) under  $n = 700$  ( $n = 2,800$ ). The ATE is also estimated by PSM using the same method as described in the application. We compare the accuracy of these semiparametric procedures to the parametric maximum likelihood (ML) and heckit two-step estimators for sample selection models. The naive estimator is also included. It simply consists of the mean difference in observed outcomes of treated and nontreated individuals and neither controls for sample selection nor treatment selection bias.

Table 5 presents the point estimates, standard deviations (st. dev.), and the mean squared errors (MSE) of the estimators. As expected, the parametric ML and two-step procedures (ML, two-step) are superior to IPW (“IPW median” for the median effect and “IPW mean” for the ATE) and PSM in terms of MSEs due to correct parametric specification. Nevertheless, the semiparametric estimators are quite competitive. The IPW estimator for the ATE and PSM even outperform the parametric methods in terms of small sample bias. In contrast, the naive estimator is severely biased.

**TABLE 5** Estimates, st. devs., and MSEs for the linear model with Gaussian errors

	$n = 700$				$n = 2800$			
	$\hat{\Delta}_{S=1}, \hat{\Delta}_{S=1}^{\tau}$	Bias	St. Dev.	MSE	$\hat{\Delta}_{S=1}, \hat{\Delta}_{S=1}^{\tau}$	Bias	St. Dev.	MSE
IPW median	0.977	-0.023	0.288	0.084	0.995	-0.005	0.145	0.021
IPW mean	0.997	-0.003	0.251	0.063	1.000	0.000	0.122	0.015
PSM	0.997	-0.003	0.252	0.064	0.999	-0.001	0.121	0.015
ML	0.982	-0.018	0.200	0.040	0.997	-0.003	0.100	0.010
Two-step	0.993	-0.007	0.214	0.046	0.997	-0.003	0.103	0.011
Naive	1.484	0.484	0.183	0.268	1.495	0.495	0.096	0.255
True	1.000				1.000			

Note: 1000 Monte Carlo replications.

We now consider the more interesting case of a nonlinear specification and treatment effect heterogeneity in  $X$ . The DGP is

$$Y_i = \alpha_1 X_i + \alpha_2 X_i^2 + \alpha_3 X_i^3 + U_i \text{ if } D_i = 1,$$

$$Y_i = \delta_1 X_i + \delta_2 X_i^2 + \delta_3 X_i^3 + U_i \text{ if } D_i = 0,$$

$$Y_i \text{ is observed if } S_i = 1,$$

$$S_i = I\{\beta_1 D_i + \beta_2 X_i + \beta_3 Z_i + V_i\} > 0,$$

$$D_i = I\{\gamma_1 X_i + \varepsilon_i\} > 0,$$

$$Y_i = Y_i^* \text{ if } S_i = 1,$$

$$X, Z \sim N(0, 1), \quad \varepsilon \sim N(0, 1), \quad U, V \sim N(0, 2), \quad \text{Cov}(U, V) = 0.8,$$

$$\text{Cov}(U, \varepsilon) = 0,$$

$$\alpha_1 = 2, \quad \alpha_2 = 6, \quad \alpha_3 = 2, \quad \delta_1 = \delta_2 = \delta_3 = 1, \quad \beta_1 = \beta_2 = 0.25,$$

$$\beta_3 = \gamma_1 = 0.5.$$

The outcome is now a cubic function of  $X$  which differs for  $D = 0, 1$ . Table 6 presents the ATE estimates for  $n = 700, 2,800$ , where all parameters are normalized with respect to the true treatment effect, such that  $\Delta_{S=1} = 1$ . In this framework with heterogenous treatment effects, IPW and PSM are strikingly more accurate than the ML and two-step estimators, which handle the nonlinearity of the outcome in  $X$  and  $D$  very poorly. In fact, their MSEs are more than 10 times larger than that of IPW under  $n = 2,800$ . This demonstrates the advantages of the more flexible semiparametric methods which invoke considerably less functional form restrictions than the standard methods applied to sample selection models.

**TABLE 6** Estimates, st. devs., and MSEs for the nonlinear model with Gaussian errors

	$n = 700$				$n = 2800$			
	$\hat{\Delta}_{S=1}$	Bias	St. Dev.	MSE	$\hat{\Delta}_{S=1}$	Bias	St. Dev.	MSE
IPW mean	1.016	0.016	0.194	0.038	1.007	0.007	0.102	0.010
PSM	1.020	0.020	0.166	0.028	1.019	0.019	0.069	0.005
ML	0.650	-0.350	0.215	0.168	0.655	-0.345	0.092	0.128
Two-step	0.665	-0.335	0.170	0.141	0.666	-0.334	0.077	0.112
Naive	1.746	0.746	0.143	0.577	1.757	0.757	0.070	0.577
True	1.000				1.000			

Note: 1000 Monte Carlo replications.

## 6. CONCLUSION

This paper discusses the identification and estimation of ATEs and QTEs in the presence of sample selection, attrition, and non-response related to unobservables. It considers the case of a double selection problem into (i) the subpopulation for which the outcome is observed (selection on unobservables) and (ii) the treatment (selection on observables). The main contribution of the paper is nonparametric identification based on weighting observations by the inverse of a nested propensity score which controls for selection bias related to being observed and being assigned to the treatment. This approach requires a continuous instrument for sample selection which needs to be — just as the treatment — conditionally independent of the unobserved factors in the model. Estimators based on IPW naturally arise from the sample analogues of the identification results. Alternatively to IPW, PSM estimators on the nested propensity score may also be applied.

In contrast to most parametric and semiparametric models, the sample selection framework considered is of rather general form. It does not require a tight specification of the relation between the selection probability, the covariates, and the outcome and allows for effect heterogeneity with respect to the observed covariates and the sample selection propensity score. Therefore, the paper shows identification of average and quantile treatment effects for various target populations, namely, the selected subpopulation (whose outcomes are observed), the entire population (irrespective of selection), and the always selected (who are selected irrespective of the treatment).

We apply IPW and PSM to U.S. labor market data previously analyzed by Mulligan and Rubinstein (2008) to determine the effect of high school graduation vs. no high school graduation on the wages of white females. The estimates suggest that graduation increases the hourly wage of working graduates and non-graduates on average by 17% in the period considered (1975 to 1979). We also estimate worst case bounds and tighter bounds based on specific assumptions concerning the sample selection process but do not obtain contradictory results.

## A. APPENDIX

### A.1. Proof of Proposition 1

Under Assumptions 1, 2, and 3,  $\Delta_{S=1}$ , the ATE on the subpopulation with observed outcomes, is identified by

$$\Delta_{S=1} = E \left[ \frac{D \cdot Y}{\pi(X, p(W))} \middle| S = 1 \right] - E \left[ \frac{(1 - D) \cdot Y}{1 - \pi(X, p(W))} \middle| S = 1 \right].$$

*Proof.*

$$\begin{aligned}
& E \left[ \frac{D \cdot Y}{\pi(X, p(W))} \middle| S = 1 \right] - E \left[ \frac{(1 - D) \cdot Y}{(1 - \pi(X, p(W)))} \middle| S = 1 \right] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ \frac{D \cdot Y}{\pi(X, p(W))} - \frac{(1 - D) \cdot Y}{(1 - \pi(X, p(W)))} \middle| X, p(W), S = 1 \right] \middle| p(W), S = 1 \right] \middle| S = 1 \right] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ \frac{Y}{\pi(X, p(W))} \middle| D = 1, X, p(W), S = 1 \right] \cdot \pi(X, p(W)) \right. \right. \\
&\quad \left. \left. - E \left[ \frac{Y}{(1 - \pi(X, p(W)))} \middle| D = 0, X, p(W), S = 1 \right] \right. \right. \\
&\quad \left. \left. \cdot (1 - \pi(X, p(W))) \middle| p(W), S = 1 \right] \middle| S = 1 \right] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ Y \middle| D = 1, X, p(W), S = 1 \right] \right. \right. \\
&\quad \left. \left. - E \left[ Y \middle| D = 0, X, p(W), S = 1 \right] \right] \middle| p(W), S = 1 \right] \middle| S = 1 \right] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ Y^1 \middle| X, p(W), S = 1 \right] \right. \right. \\
&\quad \left. \left. - E \left[ Y^0 \middle| X, p(W), S = 1 \right] \right] \middle| p(W), S = 1 \right] \middle| S = 1 \right] \\
&= E_{p(W)} \left[ E_X \left[ \Delta_{S=1}(X, p(W)) \middle| p(W), S = 1 \right] \middle| S = 1 \right] = \Delta_{S=1}.
\end{aligned}$$

The first equality follows from the law of iterated expectations, and the fourth from Assumptions 1 and 2.  $\Delta_{S=1}(X, p(W))$  denotes the conditional ATE given  $X$  and  $p(W)$  in the selected subpopulation. Finally, the last equality is a backward application of the law of iterated expectations.

## A.2. Proof of Proposition 2

Under Assumptions 1, 2, and 3,  $Q_{Y^1|S=1}^\tau$ , the  $\tau$ th quantile of  $Y^1 | S = 1$ , is an implicit function of the following expression:

$$E \left[ \frac{D}{\pi(X, p(W))} \cdot I\{Y \leq Q_{Y^1|S=1}^\tau\} \middle| S = 1 \right] = F_{Y^1|S=1}(Q_{Y^1|S=1}^\tau) = \tau.$$

*Proof.*

$$\begin{aligned}
& E \left[ \frac{D}{\pi(X, p(W))} \cdot I\{Y \leq Q_{Y^1|S=1}^\tau\} \middle| S = 1 \right] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ \frac{D}{\pi(X, p(W))} \cdot I\{Y \leq Q_{Y^1|S=1}^\tau\} \middle| X, p(W), S = 1 \right] \right] \right. \\
&\quad \left. p(W), S = 1 \right] \middle| S = 1 \Bigg] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ \frac{I\{Y \leq Q_{Y^1|S=1}^\tau\}}{\pi(X, p(W))} \middle| D = 1, X, p(W), S = 1 \right] \right. \right. \\
&\quad \left. \left. \cdot \pi(X, p(W)) \middle| p(W), S = 1 \right] \right] \middle| S = 1 \Bigg] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ I\{Y \leq Q_{Y^1|S=1}^\tau\} \middle| D = 1, X, p(W), S = 1 \right] \right. \right. \\
&\quad \left. \left. p(W), S = 1 \right] \right] \middle| S = 1 \Bigg] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ I\{Y^1 \leq Q_{Y^1|S=1}^\tau\} \middle| X, p(W), S = 1 \right] \right] \right. \\
&\quad \left. p(W), S = 1 \right] \middle| S = 1 \Bigg] \\
&= E \left[ I\{Y^1 \leq Q_{Y^1|S=1}^\tau\} \middle| S = 1 \right] = \tau.
\end{aligned}$$

The first equality follows from the law of iterated expectations, the fourth from Assumptions 1 and 2. The fifth equality is a backward application of the law of iterated expectations. An equivalent result holds for  $Q_{Y^0|S=1}^\tau$ . Therefore, the QTE  $\Delta_{S=1}^\tau$  is identified by  $Q_{Y^1|S=1}^\tau - Q_{Y^0|S=1}^\tau$ .

### A.3. Proof of Proposition 3

Under Assumptions 1, 2, 4, and 5,  $\Delta$ , the ATE on the total population, is identified by

$$\Delta = E \left[ \frac{S \cdot D \cdot Y}{p(W) \cdot \pi(X, p(W))} \right] - E \left[ \frac{S \cdot (1 - D) \cdot Y}{p(W) \cdot (1 - \pi(X, p(W)))} \right].$$

*Proof.*

$$\begin{aligned}
& E \left[ \frac{S \cdot D \cdot Y}{p(W) \cdot \pi(X, p(W))} \right] - E \left[ \frac{S \cdot (1 - D) \cdot Y}{p(W) \cdot (1 - \pi(X, p(W)))} \right] \\
&= E_{p(W)} \left[ E_X \left[ E \left[ \frac{S \cdot D \cdot Y}{p(W) \cdot \pi(X, p(W))} \right] \right. \right. \\
&\quad \left. \left. - \frac{S \cdot (1 - D) \cdot Y}{p(W) \cdot (1 - \pi(X, p(W)))} \right] \right]
\end{aligned}$$

$$\begin{aligned}
& - \frac{S \cdot (1 - D) \cdot Y}{p(W) \cdot (1 - \pi(X, p(W)))} \Big| X, p(W) \Big] p(W) \Big] \\
& = E_{p(W)} \left[ E_X \left[ E \left[ \frac{D \cdot Y}{p(W) \cdot \pi(X, p(W))} \right. \right. \right. \\
& \quad \left. \left. \left. - \frac{(1 - D) \cdot Y}{p(W) \cdot (1 - \pi(X, p(W)))} \right| S = 1, X, p(W) \right] \cdot p(W) \Big| p(W) \right] \right] \\
& = E_{p(W)} \left[ E_X \left[ E \left[ \frac{D \cdot Y}{\pi(X, p(W))} \right. \right. \right. \\
& \quad \left. \left. \left. - \frac{(1 - D) \cdot Y}{(1 - \pi(X, p(W)))} \right| S = 1, X, p(W) \right] \Big| p(W) \right] \right] \\
& = E_{p(W)} \left[ E_X \left[ E \left[ \frac{Y}{\pi(X, p(W))} \right| D = 1, S = 1, X, p(W) \right] \cdot \pi(X, p(W)) \right. \right. \\
& \quad \left. \left. - E \left[ \frac{Y}{(1 - \pi(X, p(W)))} \right| D = 0, S = 1, X, p(W) \right] \right. \right. \\
& \quad \left. \left. \cdot (1 - \pi(X, p(W))) \right| p(W) \right] \right] \\
& = E_{p(W)} \left[ E_X \left[ E \left[ Y \right| D = 1, S = 1, X, p(W) \right] \right. \right. \\
& \quad \left. \left. - E \left[ Y \right| D = 0, S = 1, X, p(W) \right] \right| p(W) \right] \right] \\
& = E_{p(W)} \left[ E_X \left[ E \left[ Y^1 \right| S = 1, X, p(W) \right] - E \left[ Y^0 \right| S = 1, X, p(W) \right] \right| p(W) \right] \right] \\
& = E_{p(W)} \left[ E_X \left[ \Delta_{S=1}(X, p(W)) \right| p(W) \right] \right] \\
& = E_{p(W)} \left[ E_X \left[ \Delta(X, p(W)) \right| p(W) \right] \right] = \Delta.
\end{aligned}$$

The first equality follows from the law of iterated expectations, the sixth from Assumptions 1 and 2. The eighth equality follows from Assumption (2.c) by which  $F_{U \mid D=d, X=x, p(W)=p(w), S=s} = F_{U \mid X=x, p(W)=p(w), S=s}$  and Assumption 5 which imposes additivity of observed and unobserved terms. Both together imply that  $\Delta_{S=1}(X, p(W))$ , the conditional ATE given  $X$  and  $p(W)$  in the selected subpopulation, is equal to  $\Delta_{S=0}(X, p(W))$  and thus,  $\Delta(X, p(W))$ . Finally, the last equality is a backward application of the law of iterated expectations.



#### A.4. Asymptotic Variance of the IPW Estimators

This section uses the results of Hirano et al. (2003), Firpo (2007), and Hahn and Ridder (2013) to derive the asymptotic variance of the IPW estimators of the ATE and QTE on the selected population under a nonparametric estimation of the selection and treatment propensity scores.

As a starting point, Hirano et al. (2003) provide the distribution of the IPW estimator of the ATE under the CIA (but in the absence of sample selection), when the treatment propensity score is estimated by series expansion and all regressors entering the propensity score are known. To this end, they invoke the following regularity conditions on the data and the propensity score model in their Assumptions 2 to 4: compact support and bounded density (also bounded away from zero) of the covariates, bounded variance of the potential outcomes, and continuous differentiability of the expected conditional outcomes given the treatment and the covariates. Furthermore, the treatment propensity score is assumed to be bounded away from zero and one and sufficiently smooth, i.e., differentiable of order  $s > 7 - r$ , where  $r$  is the dimension of the covariates. Finally, their Assumption 5 concerns the estimation of the treatment propensity score and specifies that the nonparametric series logit estimator uses a power series with the order equal to  $n^v$  for some  $1/(4(s/r - 1)) < v < 1/9$ .

Under these conditions, Theorem 1 in Hirano et al. (2003) establishes  $\sqrt{n}$ -consistency and asymptotic normality of the IPW estimator. Applied to our sample selection framework where outcomes are only observed conditional on  $S = 1$ , their results imply the following asymptotic variance ( $\text{as.Var}_{(2\text{-step})}$ ) of the estimator of the ATE on the selected population ( $\hat{\Delta}_{S=1}$ ):

$$\text{as.Var}_{(2\text{-step})} = E[(\phi + \alpha)^2 | S = 1], \quad (\text{A.1})$$

with

$$\begin{aligned} \phi &= \frac{y \cdot d}{\pi(x, p(w))} - \frac{y \cdot (1 - d)}{1 - \pi(x, p(w))} - \Delta_{S=1}, \\ \alpha &= \left( - \frac{E[Y | D = 1, X = x, p(W) = p(w)]}{\pi(x, p(w))} \right. \\ &\quad \left. - \frac{E[Y | D = 0, X = x, p(W) = p(w)]}{1 - \pi(x, p(w))} \right) \\ &\quad \times (d - \pi(x, p(w))). \end{aligned}$$

That is, there exist two terms that contribute to the asymptotic variance of  $\hat{\Delta}_{S=1}$ :  $\phi$  captures the uncertainty due to estimation of the treatment effect,

while  $\alpha$  accounts for the fact that the propensity score is not known, but has to be estimated.

An equivalent result can be derived for  $\hat{\Delta}_{S=1}^\tau$ , given that Assumptions 2, 1.A, and 2.A of Firpo (2007) are satisfied. These assumptions are closely related to Assumptions 2–5 in Hirano et al. (2003), with the exception that the potential outcomes need to be continuously distributed with unique quantiles and that their conditional distributions (given the covariates) have to be continuously differentiable in the covariates. If these restrictions hold, it follows from Theorem 1 in Firpo (2007) that the variance of  $\hat{\Delta}_{S=1}^\tau$  is of the same form as in Eq. (A.1), however, with  $\phi$  and  $\alpha$  defined in the following way:

$$\begin{aligned}\phi &= -\frac{I\{y < Q_{Y^1|S=1}^\tau\} - \tau}{f(Q_{Y^1|S=1}^\tau)} \cdot \frac{d}{\pi(x, p(w))} + \frac{I\{y < Q_{Y^0|S=1}^\tau\} - \tau}{f(Q_{Y^0|S=1}^\tau)} \cdot \frac{(1-d)}{1 - \pi(x, p(w))}, \\ \alpha &= \left( -\frac{E\left[\frac{I\{y < Q_{Y^1|S=1}^\tau\} - \tau}{f(Q_{Y^1|S=1}^\tau)} \middle| D=1, X=x, p(W)=p(w)\right]}{\pi(x, p(w))} \right. \\ &\quad \left. - \frac{E\left[\frac{I\{y < Q_{Y^0|S=1}^\tau\} - \tau}{f(Q_{Y^0|S=1}^\tau)} \middle| D=0, X=x, p(W)=p(w)\right]}{1 - \pi(x, p(w))} \right) \cdot (d - \pi(x, p(w))),\end{aligned}$$

where  $f$  denotes the pdf.

So far, we have neglected the fact that  $p(W)$ , which serves as one of the regressors in the treatment propensity score, is unknown and has to be estimated. That is, in contrast to Hirano et al. (2003) and Firpo (2007), we face a 3-step estimation problem of the kind discussed in Hahn and Ridder (2013): Estimation of (a) the selection propensity score, (b) the treatment propensity score, and (c) the effect of interest. In their Theorem 3, Hahn and Ridder (2013) provide the contribution of the nonparametric estimation of the first step to the asymptotic variance of  $\hat{\Delta}_{S=1}$  or  $\hat{\Delta}_{S=1}^\tau$ , respectively,

$$\eta = E\left[\frac{\partial^2 \phi}{\partial \pi(x, p(w))^2} \cdot (d - \pi(x, p(w))) \cdot \frac{\partial \pi(x, p(w))}{\partial p(w)} \middle| W=w\right] \cdot (s - p(w)).$$

That is, the asymptotic variance accounting for all estimation steps equals

$$\text{as.Var}_{(3\text{-step})} = E[(\phi + \alpha + \eta)^2 | S=1].$$

In empirical problems, several parameters in the variance formula (e.g., the conditional expectation  $E[Y | D = d, X = x, p(W) = p(w)]$  or the nonparametric derivative  $\frac{\partial \pi(x, p(w))}{\partial p(w)}$ ) may be difficult to estimate. Therefore, the bootstrap is likely to be an attractive alternative for inference in many applications.

## ACKNOWLEDGMENTS

I have benefited from comments by Joshua D. Angrist, Eva Deuchert, Markus Frölich, Michael Lechner, Giovanni Mellace, Blaise Melly, and Rudi Stracke, seminar/conference participants in St. Gallen, Engelberg, Bern, Innsbruck, Boston, and Mannheim, and two anonymous referees.

## REFERENCES

- Abadie, A., Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* 74:235–267.
- Ahn, H., Powell, J. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58:3–29.
- Andrews, D., Schafgans, M. (1998). Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies* 65:497–517.
- Angrist, J., Bettinger, E., Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review* 96:847–862.
- Angrist, J., Lang, D., Oreopoulos, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *American Economic Journal: Applied Economics* 1:136–163.
- Bang, H., Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* 61:962–972.
- Blundell, R., Powell, J. (2003). Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont, L. H. M., Turnovsky, S., eds. *Advances in Economics and Econometrics*. Cambridge: Cambridge University Press, pp. 312–357.
- Busso, M., DiNardo, J., McCrary, J. (2009). New evidence on the finite sample properties of propensity score matching and reweighting estimators. IZA Discussion Paper No. 3998.
- Das, M., Newey, W. K., Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies* 70:33–58.
- Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica* 75:259–276.
- Fitzgerald, J., Gottschalk, P., Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan panel study of income dynamics. *The Journal of Human Resources* 33:251–299.
- Frölich, M., Melly, B. (2008). Unconditional quantile treatment effects under endogeneity.
- Greene, W. H. (2003). *Econometric Analysis*. Upper Saddle River, NJ: Pearson Education.
- Gronau, R. (1974). Wage comparisons—a selectivity bias. *Journal of Political Economy* 82:1119–1143.
- Hahn, J., Ridder, G. (2013). The asymptotic variance of semi-parametric estimators with generated regressors. *Econometrica* 81:315–340.
- Heckman, J. J. (1974). Shadow prices, market wages and labor supply. *Econometrica* 42:679–694.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5:475–492.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* 47:153–161.
- Heckman, J. J. (1990). Varieties of selection bias. *American Economic Review, Papers and Proceedings* 80:313–318.

- Heckman, J. J., Ichimura, H., Todd, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies* 64:605–654.
- Hirano, K., Imbens, G. W., Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71:1161–1189.
- Horvitz, D. G., Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47:663–685.
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* 87:706–710.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* 86:4–29.
- Imbens, G. W., Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* 62:467–475.
- Imbens, G. W., Newey, W. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* 77:1481–1512.
- Khan, S., Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica* 78:2021–2042.
- Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica* 46:33–50.
- Lechner, M. (1999). Earnings and employment effects of continuous off-the-job training in east germany after unification. *Journal of Business and Economic Statistics* 17:74–90.
- Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In: Lechner, M., Pfeiffer, F., eds. *Econometric Evaluations of Active Labor Market Policies in Europe*. Heidelberg: Physica.
- Lechner, M. (2007). A note on the relation of weighting and matching estimators. University of St. Gallen Discussion Paper no. 2007-34.
- Lechner, M., Melly, B. (2007). Earnings effects of training programs. IZA Discussion Paper no. 2926.
- Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies* 76:1071–1102.
- Manski, C. F. (1989). Anatomy of the selection problem. *The Journal of Human Resources* 24:343–360.
- Manski, C. F. (1994). The selection problem. In: Sims, C., ed. *Advances in Econometrics: Sixth World Congress*. Cambridge: Cambridge University Press, pp. 143–170.
- Martins, M. F. O. (2001). Parametric and semiparametric estimation of sample selection models: An empirical application to the female labour force in portugal. *Journal of Applied Econometrics* 16:23–39.
- Meng, C.-L., Schmidt, P. (1985). On the cost of partial observability in the bivariate probit model. *International Economic Review* 26:71–85.
- Mroz, T. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55:765–799.
- Mulligan, C. B., Rubinstein, Y. (2008). Selection, investment, and women's relative wages over time. *Quarterly Journal of Economics* 123:1061–1110.
- Newey, W., Powell, J., Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica* 67:565–603.
- Newey, W. K. (2007). Nonparametric continuous/discrete choice models. *International Economic Review* 48:1429–1439.
- Newey, W. K. (2009). Two-step series estimation of sample selection models. *Econometrics Journal*.
- Newey, W. K., Powell, J. L., Walker, J. (1990). Semiparametric estimation of selection models: Some empirical results. *American Economic Review* 80:324–328.
- Poirier, D. J. (1980). Partial observability in bivariate probit models. *Journal of Econometrics* 12: 209–217.
- Politis, D. N., Romano, J. P., Wolf, M. (1999). *Subsampling*. New York: Springer-Verlag.
- Robins, J., Rotnitzky, A., Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of American Statistical Association* 90:106–121.
- Robins, J., Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association* 90:122–129.
- Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.
- Rotnitzky, A., Robins, J. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* 82:805–820.

- Rubin, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics* 29:159–183.
- Rubin, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 29:185–203.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688–701.
- Rubin, D. B. (1976). Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics* 32:109–120.
- Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* 25:279–292.
- Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software* 42:1–52.
- Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources* 33:127–169.
- Wooldridge, J. (2002). Inverse probability weighted m-estimators for sample selection, attrition and stratification. *Portuguese Economic Journal* 1:141–162.
- Wooldridge, J. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics* 141:1281–1301.
- Zhang, J., Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcome are truncated by death. *Journal of Educational and Behavioral Statistics* 28:353–368.
- Zhang, J., Rubin, D. B., Mealli, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. In: Millimet, D., Smith, J., Vytlačil, E., eds. *Advances in Econometrics: Modelling and Evaluating Treatment Effects in Econometrics*, vol. 21, Amsterdam: Elsevier Science Ltd., pp. 117–145.